Eur Arch Psychiatry Clin Neurosci (2004) 254 : 215–223 DOI 10.1007/s00406-004-0476-3

ORIGINAL PAPER

Verena Henkel · Roland Mergl · James C. Coyne · Ralf Kohnen · Hans-Jürgen Möller · Ulrich Hegerl

Screening for depression in primary care: Will one or two items suffice?

Received: 22 October 2003 / Accepted: 18 November 2003

Abstract Small differences in implementation of screening and the associated burden on clinicians and patients could have substantial effects on the sustainability of screening in routine primary care. Therefore, we investigated the psychometric properties of single items and two-item combinations of the "WHO-5 Well Being Index" (WHO-5) and compared the obtained characteristics to those of the original version as well as to another proposed two-item screener (developed from PRIME-MD and BPHQ, respectively).

Screening and diagnostic interview data from 431 primary care patients were analysed. Main outcome measures were sensitivity, specificity and AUC values. All test characteristics were assessed using the diagnoses derived from the Composite International Diagnostic Interview (CIDI) as the criterion standard.

Single-item screening questions proved rather inadequate. However, only marginal differences in performance were found between two questions and the longer screening instrument with respect to major depression, dysthymia and "any depressive disorder". There were no statistically significant differences between these AUC values and most other test characteristics assessed.

The results suggest that screening could be reduced to two questions with a potential advantage in terms of ease of administration and scoring and decreased staff and patient burden and perhaps a reduced stigma associated with a positive screening score.

Key words depression · primary care · screening · brief screeners

Dr. Verena Henkel, M. D. (✉) · R. Mergl, M.Sc. · H.-J. Möller, Prof., M. D. · U. Hegerl, Prof., M. D.
Department of Psychiatry
Ludwig-Maximilians-University
Nußbaumstr. 7
80336 Munich, Germany
Tel.: +49-89/5160-5558
Fax: +49-89/5160-5542
E-Mail: verena.henkel@psy.med.uni-muenchen.de

Prof. J. C. Coyne, Ph. D.
Department of Psychiatry
University of Pennsylvania Health System
11 Gates, 3400 Spruce St
PA-19104 Philadelphia, USA

Prof. R. Kohnen, Ph. D.
Institute for Medical Research Management and Biometrics (IMEREM)
Scheurlstr. 21
90478 Nuremberg, Germany

Introduction

Depressive disorders are widely distributed in the general population [17, 20] and are among the most common conditions in primary care settings [22], but adequate recognition and accurate diagnosis have proven frustratingly difficult to achieve in routine care. Although there is evidence that rates of detection and treatment have improved over the last decade [17], many depressed persons in the community remain undetected as well as untreated [6, 8, 22] and primary care is a critical context for identifying them. There is evidence that almost half of primary care patients with current major depression will at some point develop suicidal ideation [27], often in periods between primary care visits, which gives increased importance to detection of depression in these visits.

Depression is a clinical diagnosis based on medical history, the description of symptoms and the exclusion of competing diagnoses. There is no specific biomarker, no physiological or laboratory test to definitively assess the diagnosis. However, two main criteria for mass screening are met: 1) the high prevalence of the disorder and 2) the ready availability of treatments with well documented efficacy and tolerability. Nevertheless, two other related important issues remain controversial: 1) which circumstances are necessary to ensure that screening be sustained and favourably influences outcome?; and 2) which screening test has the optimal bal-

ance of good operating characteristics and potential for wide acceptance by patients as well as physicians? The present study addresses the latter issue as a way of approaching the former. A number of studies have already suggested use of self-rated screening questionnaires followed by a clinical interview with patients screening positive in order to improve recognition of depression (for review see [34]). Results of these studies have demonstrated that various depression screeners have reasonable, similar test characteristics, so that selection of the instrument can focus on feasibility, administration and scoring times as well as patients' acceptance [33, 34]. Small differences in the demands of screening on clinical staff and patients can have potential substantial effects on the ability to sustain screening in routine care and ensure its effectiveness in terms of improved patient outcomes. Thus, a review showed that the effectiveness of screening depended on physicians not having to contend with information that patients had screened negative [12].

In a previous study our group had demonstrated that the "WHO-5 Well Being Index" (WHO-5) is a brief and highly sensitive screening tool (sensitivity: 93 %; specificity: 64 %) [14]. However, it has proven difficult to get physicians and staff to comply with the demands of screening and systematically following up of positively screening patients [19, 24, 29], even when physician prompts and additional staff support are provided [31]. Anything that reduces these barriers (such as reducing screening to fewer questions) could potentially have payoffs for the prospect of implementing and sustaining screening within the competing demands of routine care. There is a view that two screening questions are sufficient [3, 13, 25, 33] (e. g. about depressed mood and anhedonia during the past month [33]). Whooley et al. [33] reported a 96 % sensitivity (specificity: 57 %) for these two items of the Primary Care Evaluation of Mental Disorders (PRIME-MD [30]). However, this study [33] was not conducted in a primary care sample. If the results of this study could be replicated in a primary care sample and if there would be only marginal differences in performance between two questions and a longer instrument, then the two-question approach could have substantial advantage in terms of ease of administration and scoring as well as reduced staff and patient burden. Going to fewer questions and reducing the need for scoring could be a step in the right direction in terms of reducing consumption of resources. Therefore, we were interested in the test characteristics of single items and of two-item combinations of the WHO-5 questionnaire in a primary care population in order to investigate if the WHO-5 screening instrument could be abbreviated. In addition, we assessed the test characteristics of the two-item screener proposed by Whooley et al. [33]. In a second step, both two-item versions and the original WHO-5 version were compared. We were especially interested in the performance of the screeners with respect to major depression and dysthymia, because established treatment guidelines based on empirical demonstration of treatment efficacy are available [10].

Patients and method

This report relies on data from a larger study comparing different methods of detection of depression in primary care [14] that involved 431 primary care patients seen by 18 primary care physicians (post hoc reanalysis) (for the description of the sample see Table 1). The study protocol was approved by the Ethics Review Committee. Written informed consent was obtained from all subjects before study start.

On days predetermined by the participating practices, all adult patients who presented routinely in the waiting room were invited to participate in the study. Before being seen by their physician, patients completed the screening instruments. Within six days of their visit, patients were contacted by phone and a fully structured, standardised psychiatric interview (CIDI) [35] was conducted. The 17 subjects who failed to keep CIDI appointments or refused participation in the interview were excluded.

The prevalence of "any depressive disorder" was 16.7 % in our study (72 of 431; 95 % C. I. 12.8–21.1 %). In this group, 43 patients suffered from a major depressive episode, 22 patients from dysthymia and 7 patients from other depressive diagnoses (always current at the time of the interview) (Table 1).

Criterion standard

The Composite International Diagnostic Interview (CIDI) is a fully structured instrument for use by trained interviewers. This instrument had been selected because reliability and validity have been established [1, 35]. We used a computer-administered form (DIA-X) [36], based on CIDI version 1.1. [37]. The equivalency of the CIDI delivered by human interviewers and its computerised version has been confirmed [23]. All interviewers (six psychologists and one psychiatrist) were trained by a designated CIDI training centre (psychiatric department, Max-Planck Institute, Munich). Thus, in contrast to the common use of lay interviewers, only trained mental health specialists administered and interpreted the CIDI in the present study.

Screening questionnaires

The "Brief Patient Health Questionnaire" (BPHQ) was developed at the end of the 1990s by Spitzer et al. [30] as an abbreviated modification of the PRIME-MD. In the present study, only the first two items (depressed mood and loss of interest) were considered following the

Table 1 Description of the sample of 431 primary care patients

	Any depressive disorder (acc. to CIDI)	No depression (acc. to CIDI)
N (%)	72 (16.7 %) – Major depression (ICD-10: F32/33): N = 43 (10 %) – Dysthymia (ICD-10: F34): N = 22 (5.1 %) – Other depressive diagnoses (ICD-10: F31, F06): N = 7 (1.6 %)	359 (83.3 %)
Age (± SD)	47.2 (± 15.3) years	53.7 (± 16.8) years
Gender (N)	49 ♀/23 ♂	223 ♀/136 ♂

suggestions by Whooley et al. [33] (see Table 2). The scoring procedure was conducted according to Spitzer et al. [30]. For interpretation of our findings, it is important to know that there is one difference between the two-item version of the PRIME-MD suggested [3, 33] and the tested version in our study. In the BPHQ all questions refer to the past two weeks, whereas the depression module of the PRIME-MD refers to the past month.

The "WHO-5 Well Being Index" (WHO-5) (see Table 2) was developed by Bech in the 1990s [15, 38]. It is a set of five items to measure the degree of well-being. Scoring was conducted as suggested by the World Health Organisation [15]. Test characteristics of each single item and of each possible two-item combination had been assessed. Among these combinations tested, the two-item combination with the highest AUC value has been selected for further comparative statistical analyses in the present study.

Data analysis

Statistical analyses were performed using the statistical software SPSS (Statistical Package for Social Sciences) for Windows (version 10.0) and SAS statistical software (Release 6). Binomial 95 % confidence intervals (C. I.) have been used in order to determine the precision of the estimates for the prevalence of depressive disorders. Receiver operating characteristic (ROC) curves from data arising from simple random sampling (for details see [11]) were constructed to describe and to visually compare the screening tools. The areas under the curves (AUC values) were established using the trapezoidal rule and SPSS (Version 10.0). 95 % C. I. for the AUC values was computed using bootstrapping methods. In a next step, it was determined, whether the AUC values were statistically different using a nonparametric method for correlated samples [9]. For this purpose, SAS statistical software (Release 6) providing a command syntax for the above-mentioned test was used.

For the WHO-5 two-item version the cut-off point was selected in a way that sensitivity approximated 90 %. For the other two screening tools (WHO-5 original version and the BPHQ two-item version) the original cut-off scores – following the corresponding literature [15, 33] – were chosen. Then, two-by-two tables were constructed, displaying always screening instrument diagnosis (positive/negative) by CIDI diagnosis of major depression, dysthymia as well as "any depressive disorder" (positive/negative). Based on these two-by-two tables, overall quality of the screening instruments was assessed using Cohen's kappa measuring the agreement of the screening tool with the reference standard (CIDI) over and above the agreement that would be expected by chance. Measures of specificity for these three screening tools were calculated including exact binomial 95 % C. I. and were compared (WHO-5 original version versus WHO-5 two-item version; WHO-5 original version versus BPHQ two-item version as well as WHO-5 two-item version versus BPHQ two-item version) using McNemar tests. For these three comparisons, two-sided tests have been chosen and the significance level has been alpha-adjusted ($\alpha = 0.017$). Moreover, positive and negative predictive values, positive likelihood ratios (sensitivity/1-specificity) as well as negative likelihood ratios (1-sensitivity/specificity) were calculated in order to verify diagnostic accuracy of the three above-mentioned screening instruments.

Results

Depression as currently diagnosed probably represents a heterogeneous set of disorders. Therefore, data analyses had been conducted with respect to different depressive diagnoses: 1) *"any depressive disorder"* including *all* depressive diagnoses in our study; 2) *major depression* and 3) *dysthymia*, because major depression and dysthymia have established guidelines based on empirical demonstrations of treatment efficacy [10].

Analysis for "any depressive disorder"

AUC values of each single item of WHO-5 are listed in Table 3. Among the single items of the WHO-5, the first item ("I have felt cheerful and in good spirits") had the

Table 2 Items of the screening instruments for depression in primary care selected in the present study

WHO-5 items
No. 1: "I have felt cheerful and in good spirits."
No. 2: "I have felt calm and relaxed."
No. 3: "I have felt active and vigorous."
No. 4: "I woke up feeling fresh and rested."
No. 5: "My daily life has been filled with things that interested me."
BPHQ two-item version: items
No. 1: "During the past two weeks, have you often been bothered by feeling down, depressed, or hopeless?"
No. 2: "During the past two weeks, have you often been bothered by little interest or pleasure in doing things?"

BPHQ Brief Patient Health Questionnaire [30]; *WHO-5* WHO-5 Well Being Index [15, 38]

Table 3 Test characteristics for each item of the WHO-5; analysis for "any depressive disorder"

WHO-5 item	Patients *with* depression (N = 72) (M ± s) (range)	Patients *without* depression (N = 359) (M ± s) (range)	Area Under ROC Curve (95 % C. I.)
No. 1: "I have felt cheerful and in good spirits."	1.43 ± 1.05 (0–5)	3.14 ± 1.18 (0–5)	0.85 (0.80–0.89)*
No. 2: "I have felt calm and relaxed."	1.31 ± 1.12 (0–5)	2.90 ± 1.39 (0–5)	0.80 (0.75–0.85)*
No. 3: "I have felt active and vigorous."	1.03 ± 1.16 (0–4)	2.83 ± 1.37 (0–5)	0.83 (0.78–0.88)*
No. 4: "I woke up feeling fresh and rested."	1.04 ± 1.11 (0–5)	2.77 ± 1.52 (0–5)	0.81 (0.76–0.86)*
No. 5: "My daily life has been filled with things that interested me."	1.69 ± 1.33 (0–5)	3.44 ± 1.32 (0–5)	0.81 (0.75–0.87)*

C. I. 95 % Confidence Interval; *M* mean; *N* sample size; *s* standard deviation; * $p \leq 0.001$ (asymptotic significance; null hypothesis: true area under ROC curve = 0.5)

highest AUC value (0.85), followed by the third item ("I have felt active and vigorous") (0.83). AUC values for the other items were only within a range of 0.80 to 0.81 for the point estimates.

A small range of AUC values (0.83–0.86) was found for the ten possible two-item combinations of the WHO-5, with the sum score computed for the combination of items 1 and 3 indicating the highest AUC value (0.86; 95 % C. I. 0.82–0.91). We considered the 95 % C. I. to be a reasonable criterion for the selection of this two-item combination of the WHO-5 for further analysis. Fig. 1 demonstrates the ROC curves for the three brief screening instruments for "any depressive disorder".

In Table 4, sensitivities, specificities, false positive rates as well as predictive values are summarised for each possible two-item combination of WHO-5 using those cut-off scores which provide comparable high values for sensitivity (in each case about 90 %), since sensitivity is the most important property of a screening instrument (no patient suffering from the disorder should be missed) [14].

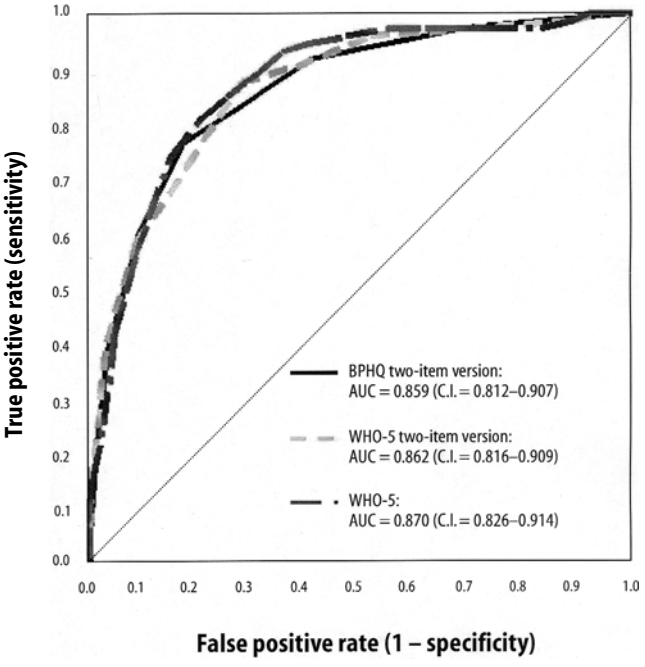

Fig. 1 Receiver operating characteristic (ROC) curves for WHO-5 (original version) as well as the two-item versions of the WHO-5 and the BPHQ for *all* subjects (for "any depressive disorder" according to CIDI). The diagonal line represents the null hypothesis (area under curve (AUC) value = 0.50). *CI* 95 % confidence interval

The ability of the two-item combination of the WHO-5 (item 1 and 3) to discriminate between individuals with or without depression did not significantly differ either from that of the complete original WHO-5 questionnaire ($\chi^2 = 0.76$; df = 1; p = 0.38) or from the two-item version of the BPHQ ($\chi^2 = 0.03$; df = 1; p = 0.86). The AUC value of the sum score computed for the original version of the WHO-5 did not significantly differ from that determined for the two-item version of the BPHQ ($\chi^2 = 0.48$; df = 1; p = 0.49). Table 5 summarises the most important operating characteristics of the above mentioned screening tools based on different cut-off points. The differences between these three instruments regarding sensitivity were not significant ($0.33 \leq \chi^2 \leq 2$; df = 1; $0.50 \leq p \leq 1$). Specificity of the *original* version of the WHO-5 was not significantly superior to specificity of the BPHQ two-item version ($\chi^2 = 4.00$; df = 1; p = 0.05), as well as not significantly superior to that of the *WHO-5 two-item version* ($\chi^2 = 2.07$; df = 1; p = 0.15). Furthermore, the WHO-5 *two-item* version and the BPHQ *two-item* version did not significantly differ regarding specificity ($\chi^2 = 0.96$; df = 1; p = 0.33). A Bonferroni correction has been applied for each of these comparative analyses.

Regarding the *predictive values*, screening 431 patients yielded 195 patients (45.2 %) screening positive in the WHO-5, and of these patients 67 patients (34.4 %) were found to have any depressive disorder upon follow-up interview. Of the 236 patients screening negative in the WHO-5 (54.8 %), 231 (97.9 %) were not depressed according to CIDI. Similarly, 202 patients (46.9 %) were screened positive in the two-item version of WHO-5, with 65 patients (32.2 %) suffering from any depressive disorder. Using this screener, 229 patients (53.1 %) were

Table 4 Test characteristics for each possible two-item version of WHO-5; analysis for "any depressive disorder"

Two-item combination of the WHO-5 Well Being Index	Cut-off score	Sensitivity (%) (95 % C. I.)	Specificity (%) (95 % C. I.)	False positive rate (%) (95 % C. I.)	Positive predictive value (%) (95 % C. I.)	Negative predictive value (%) (95 % C. I.)
Item No. 1 + Item No. 2	≤ 5	88.9 (79.3–95.1)	63.5 (58.3–68.5)	36.5 (31.5–41.7)	32.8 (26.3–39.9)	96.6 (93.4–98.5)
Item No. 1 + Item No. 3	≤ 5	90.3 (81.0–96.0)	61.8 (56.6–66.9)	38.2 (33.1–43.4)	32.2 (25.8–39.1)	96.9 (93.8–98.8)
Item No. 1 + Item No. 4	≤ 4	90.3 (81.0–96.0)	70.8 (65.8–75.4)	29.3 (24.6–34.3)	38.2 (30.9–46)	97.3 (94.6–98.9)
Item No. 1 + Item No. 5	≤ 5	87.5 (77.6–94.1)	72.4 (67.5–77.0)	27.6 (23.0–32.5)	38.9 (31.3–46.9)	96.7 (93.7–98.5)
Item No. 2 + Item No. 3	≤ 5	91.7 (82.7–96.9)	57.7 (52.4–62.8)	42.3 (37.2–47.6)	30.3 (24.3–36.8)	97.2 (94–99)
Item No. 2 + Item No. 4	≤ 4	90.3 (81.0–96.0)	67.1 (62.0–72.0)	32.9 (28.0–38.0)	35.5 (28.6–42.9)	97.2 (94.3–98.9)
Item No. 2 + Item No. 5	≤ 5	86.1 (75.9–93.1)	68.8 (63.7–73.6)	31.2 (26.4–36.3)	35.6 (28.5–43.2)	96.1 (93–98.1)
Item No. 3 + Item No. 4	≤ 4	87.5 (77.6–94.1)	65.5 (60.3–70.4)	34.5 (29.6–39.7)	33.7 (27–40.9)	96.3 (93.1–98.3)
Item No. 3 + Item No. 5	≤ 5	88.9 (79.3–95.1)	66.0 (60.9–70.9)	34.0 (29.1–39.1)	34.4 (27.6–41.7)	96.7 (93.7–98.6)
Item No. 4 + Item No. 5	≤ 5	88.9 (79.3–95.1)	63.5 (58.3–68.5)	36.5 (31.5–41.7)	32.8 (26.3–39.9)	96.6 (93.4–98.5)

C. I. 95 % Confidence Interval

Table 5 Operating characteristics for three depression case-finding instruments based on cut-off points following the literature (except for the WHO-5 two-item version); analysis for "any depressive disorder"

Instrument	Sensitivity, % (95 % C. I.)	Specificity, % (95 % C. I.)	Positive predictive value, % (95 % C. I.)	Negative predictive value, % (95 % C. I.)	False positive rate (95 % C. I.)	Kappa (95 % C. I.)	Likelihood Ratio	
							Positive	Negative
WHO-5 original version (cut-off score ≤ 13)	93.1 (84.5–97.7)	64.4 (59.2–69.3)	34.4 (27.7–41.5)	97.9 (95.1–99.3)	35.6 (30.7–40.9)	0.34 (p ≤ 0.001) (0.27–0.42)	2.61	0.11
WHO-5 two-item version (item 1 + 3) (cut-off score ≤ 5)	90.3 (81–96.0)	61.8 (56.6–66.9)	32.2 (25.8–39.1)	96.9 (93.8–98.8)	38.2 (33.1–43.4)	0.30 (p ≤ 0.001) (0.23–0.38)	2.37	0.16
BPHQ two-item version (cut-off score ≥ 4)	91.7 (82.7–96.9)	59.1 (53.8–64.2)	31 (24.9–37.7)	97.3 (94.1–99)	40.9 (35.8–46.2)	0.28 (p ≤ 0.001) (0.21–0.36)	2.24	0.14

C. I. Confidence Interval

screened negative; of these patients, 222 patients (96.9 %) were found to have no depressive disorder. Moreover, screening 431 patients yielded 213 patients (49.4 %) screening positive in the two-item version of the BPHQ; of these patients, only 66 patients (31 %) had a CIDI diagnosis of any depressive disorder. In contrast, 218 patients (50.6 %) were screened negative in this two-item screener and 212 of these patients (97.3 %) did not suffer from depression. Since predictive values vary according to prevalence, we gave more emphasis to other operating characteristics which do not vary according to prevalence. The kappa coefficients of the three depression case-finding instruments provided similar values (0.28–0.34) all indicating that their agreement with the reference standard was significantly above the agreement that would be expected by chance ($p \leq 0.001$). All three screening tools were characterised by good negative likelihood ratios (0.11–0.16) and modest positive likelihood ratios (2.24–2.61).

Analysis for major depression

AUC values of each possible two-item combination of WHO-5 were nearly equivalent with a range of 0.83 to 0.85, with the sum score computed for the combination of item 1 and item 4 indicating the highest AUC value (0.85; 95 % C. I. 0.79–0.91). This AUC value is nearly identical with the AUC value computed for the original version of the WHO-5 (0.86; 95 % C. I. 0.81–0.91) and the BPHQ two-item version (0.86; 95 % C. I. 0.81–0.92).

In Table 6 the test characteristics are summarised for each possible two-item combination of WHO-5 using those cut-off scores with comparable high values for sensitivity (about 90 %). The differences between the original WHO-5 questionnaire, the WHO-5 two-item version (item 1 and item 4) and the BPHQ two-item version regarding sensitivity were not statistically significant. Moreover, specificity of the original version of the WHO-5 was not significantly superior to that of the BPHQ two-item version ($\chi^2 = 4.00$; df = 1; p = 0.05), but significantly lower than that of the WHO-5 two-item version ($\chi^2 = 15.61$; df = 1; $p \leq 0.001$). The WHO-5 two-item version also demonstrated a significantly higher

Table 6 Test characteristics for each possible two-item version of WHO-5; analysis for major depression (ICD-10: F32/33)

Two-item combination of the WHO-5 Well Being Index	Cut-off score	Sensitivity (%) (95 % C. I.)	Specificity (%) (95 % C. I.)	False positive rate (%) (95 % C. I.)	Positive predictive value (%) (95 % C. I.)	Negative predictive value (%) (95 % C. I.)
Item No. 1 + Item No. 2	≤ 5	86.1 (72.1–94.7)	63.5 (58.3–68.5)	36.5 (31.5–41.7)	22.0 (16.0–29.1)	97.4 (94.5–99.1)
Item No. 1 + Item No. 3	≤ 5	90.7 (67.9–97.4)	61.8 (56.6–66.9)	38.2 (33.1–43.4)	22.2 (16.3–29.0)	98.2 (95.5–99.5)
Item No. 1 + Item No. 4	≤ 4	88.4 (74.9–96.1)	70.8 (65.8–75.4)	29.3 (24.6–34.3)	26.6 (19.5–34.6)	98.0 (95.6–99.4)
Item No. 1 + Item No. 5	≤ 6	90.7 (77.9–97.4)	57.4 (52.0–62.6)	42.6 (37.4–48.0)	20.3 (14.9–27.0)	98.1 (95.2–99.5)
Item No. 2 + Item No. 3	≤ 5	90.7 (77.9–97.4)	57.7 (52.4–62.8)	42.3 (37.2–47.6)	20.4 (14.9–26.8)	98.1 (95.2–99.5)
Item No. 2 + Item No. 4	≤ 4	90.7 (77.9–97.4)	67.1 (62.0–72.0)	32.9 (28.0–38.0)	24.8 (18.3–32.4)	98.4 (95.9–99.6)
Item No. 2 + Item No. 5	≤ 6	93.0 (81–99)	51.8 (46.5–57.1)	48.2 (42.9–53.5)	18.8 (13.8–24.7)	98.4 (95.4–99.7)
Item No. 3 + Item No. 4	≤ 4	86.1 (72.1–94.7)	65.5 (60.3–70.4)	34.5 (29.6–39.7)	23.0 (16.7–30.3)	97.5 (94.7–99.1)
Item No. 3 + Item No. 5	≤ 6	90.7 (77.9–97.4)	54.0 (48.7–59.3)	46.0 (40.7–51.3)	19.1 (13.7–25.2)	98 (95–99.5)
Item No. 4 + Item No. 5	≤ 6	93.0 (80.9–98.4)	49.6 (44.3–54.9)	50.4 (45.1–55.7)	18.1 (13.3–23.9)	98.3 (95.2–99.7)

C. I. 95 % Confidence Interval

specificity than the BPHQ two-item version ($\chi^2 = 19.10$; df = 1; $p \leq 0.001$) (always after Bonferroni correction).

Analysis for dysthymia

Regarding the AUC values, the range was small, with the sum score computed for the combination of item 1 and item 3 indicating the highest AUC value (0.91; 95 % C. I. 0.86–0.96). This AUC value is even better than that for the original version of the WHO-5 (0.87; 95 % C. I. 0.82–0.93) and the BPHQ two-item version (0.87; 95 % C. I. 0.80–0.94).

In Table 7, test characteristics are summarised for each possible two-item combination of WHO-5 selecting those cut-off scores which provide comparable high values for sensitivity (in each case about 90 %). Performances of the original WHO-5, the WHO-5 two-item version (item 1 and item 3) and the BPHQ two-item version regarding sensitivity were comparable. However, using specified cut-off scores, specificity of the *original* version of the WHO-5 was significantly superior to that of the BPHQ two-item version ($\chi^2 = 55.74$; df = 1; $p \leq 0.001$) as well as to that of the WHO-5 two-item version (item 1 and item 3) ($\chi^2 = 17.33$; df = 1; $p \leq 0.001$). The WHO-5 two-item version also demonstrated a significantly higher specificity than the BPHQ *two-item* version ($\chi^2 = 25.96$; df = 1; $p \leq 0.001$) (always after Bonferroni correction).

Comment

There were two major sequential objectives of this study:

1. To investigate if the diagnostic validity of single items or of two-item combinations from the "WHO-5 Well Being Index" [15] (WHO-5) would be similar to the diagnostic performance of the longer original version; and
2. To compare the operating characteristics of the obtained short version of WHO-5 as well as of the original version of WHO-5 to those of the previously suggested two-item version of the PRIME-MD [33] (BPHQ [30] respectively).

Objective 1. Although the performance of single items was rather inadequate, our results are in line with the findings of other studies [3, 25, 33] and of recently published recommendations [13] indicating that screening can be as simple as asking two questions. Even a two-item questionnaire appears to be sufficient as the first stage in a two-stage process of screening and interview follow-up of patients who screened positive.

Only marginal differences were found for the performance for each possible two-item combination of the WHO-5, making the selection of the "best" two-item combination somewhat arbitrary. If the 95 % C. I. of the AUC values of "any depressive disorder" was considered to be a reasonable criterion for this choice, then the combination of a psychological feature (item 1: "I have felt cheerful and in good spirits") and a more somatic feature (item 3: "I have felt active and vigorous") appeared to be the optimal two-item combination of the WHO-5, at least for the group of all depressive diagnoses ("any depressive disorder") and for dysthymia.

However, if special cut-off scores providing a reasonable balance between sensitivity and specificity were considered, another two-item combination of the WHO-5 would be favoured. Interestingly, this combination also connects a psychological feature (item 1 ("mood"): "I have felt cheerful and in good spirits") with a more somatic feature (item 4 ("sleep"): "I woke up feeling fresh and rested."). Applying a cut-off score of 4, for this two-item combination a quite high sensitivity (90.3 %) and a moderate specificity (70.8 %) could be computed in our sample for "any depressive disorder". For patients with "major depression" it was now this item combination (1 + 4) which also performed best in terms of the AUC value (in contrast, the AUC value of items 1 + 3 performed best in patients with "dysthymia" or "any depressive disorder") probably reflecting differences in

Table 7 Test characteristics for each possible two-item version of WHO-5; analysis for dysthymia (ICD-10: F34)

Two-item combination of the WHO-5 Well Being Index	Cut-off score	Sensitivity (%) (95 % C. I.)	Specificity (%) (95 % C. I.)	False positive rate (%) (95 % C. I.)	Positive predictive value (%) (95 % C. I.)	Negative predictive value (%) (95 % C. I.)
Item No. 1 + Item No. 2	≤ 5	95.5 (77.2–99.9)	63.5 (58.3–68.5)	36.5 (31.5–41.7)	13.8 (8.8–20.3)	99.6 (97.6–100)
Item No. 1 + Item No. 3	≤ 4	95.5 (77.2–99.9)	71.9 (66.9–76.5)	28.1 (23.4–33.1)	17.2 (11–25.1)	99.6 (97.9–100)
Item No. 1 + Item No. 4	≤ 4	95.5 (77.2–99.9)	70.8 (65.8–75.4)	29.3 (24.6–34.3)	16.7 (10.6–24.3)	99.6 (97.8–100)
Item No. 1 + Item No. 5	≤ 5	95.5 (77.2–99.9)	72.4 (67.5–77)	27.6 (23.0–32.5)	17.5 (11.2–25.5)	99.6 (97.9–100)
Item No. 2 + Item No. 3	≤ 5	95.5 (77.2–99.9)	57.7 (52.4–62.8)	42.3 (37.2–47.6)	12.1 (7.7–17.8)	99.5 (97.4–100)
Item No. 2 + Item No. 4	≤ 3	90.9 (70.8–98.9)	75.8 (71–80.1)	24.2 (19.9–29.0)	18.7 (11.8–27.4)	99.3 (97.4–99.9)
Item No. 2 + Item No. 5	≤ 5	90.9 (70.8–98.9)	68.8 (63.7–73.6)	31.2 (26.4–36.3)	15.2 (9.5–22.4)	99.2 (97.1–99.9)
Item No. 3 + Item No. 4	≤ 3	95.5 (77.2–99.9)	74.7 (69.8–79.1)	25.3 (20.9–30.2)	18.8 (12.0–27.2)	99.6 (98–100)
Item No. 3 + Item No. 5	≤ 5	95.5 (77.2–99.9)	66.0 (60.9–70.9)	34 (29.1–39.1)	14.7 (9.3–21.6)	99.6 (97.7–100)
Item No. 4 + Item No. 5	≤ 5	95.5 (77.2–99.9)	63.5 (58.3–68.5)	36.5 (31.5–41.7)	13.8 (8.8–20.3)	99.6 (97.6–100)

C. I. 95 % Confidence Interval

the psychopathology of the different subtypes of depression.

In view of the fact that the combination of items 1 and 3 performed best in the whole group of "any depressive disorder" as well as for "dysthymia" in terms of AUC values and considering ROC analysis to be superior to the traditional approach using specified cut-off scores [11], we have selected this two-item combination (WHO-5 item 1 + 3) for further statistical comparisons (second objective). Since WHO-5 items 1 and 4 performed best in major depression, we selected this combination for further comparative analysis in this category.

Objective 2. There were basically only marginal differences in test characteristics between all three screeners across the diagnostic categories. No significant differences were found for the comparisons between WHO-5 original version and both two-item versions as well as between the two-item version of WHO-5 (item 1 + 3) and the BPHQ two-item version for "any depressive disorder". For the analysis of "dysthymia" there were significant results which sustained after Bonferroni correction in favour of the WHO-5 original version compared to both two-item screeners. Interestingly, the two-item version of WHO-5 (item 1 + 3) was statistically superior to the two-item version of BPHQ. For the analysis of "major depression" the two-item version of WHO-5 (item 1 + 4) was now superior to the original version of WHO-5 as well as to the BPHQ two-item version. However, these differences in the analyses for "dysthymia" and "major depression" were not obvious in the ROC analyses.

Our key substantive findings are that two screening items work as well as more items and that patients' failure to endorse fully two items that indicate they are feeling cheerful and energetic is at least as indicative of likelihood of a depressive disorder as their endorsement of the cardinal symptoms of depressed mood and anhedonia. We found no basis for recommending longer screening instruments over two items. Moreover, an advantage of needing only two items is that they can readily be visually inspected without a formal scoring and it may be well that a separate screening instrument can be dispensed with altogether if the two questions were added to the usual clinic intake form or verbally asked by clinical staff. It has proposed in the palliative care literature that simply verbally asking a couple of questions may be as effective in identifying depressed patients as more formal screening procedures [4], and our data suggest this issue may be worthy of attention in primary care settings.

We find it noteworthy that positive items addressing cheerfulness and energy level work as well for the purposes of screening as narrowly symptom-oriented items. This is despite normative data [16, 28] and conventional wisdom [2] suggesting positive and negative mood are so distinct enough that the absence of positive well being is not indicative of negative well-being. Perhaps these previous observations do not apply to risk of clinical disorder, but, regardless, there may be practical advantages in focusing brief screening on the absence of positive well-being, rather than the presence of formal symptoms. Namely, patients who find conventional screening intrusive or stigmatising might respond more favourably to inquiries about positive well-being. This issue may be relevant to making inroads in the approximately 10–40 % of primary care patients who reject opportunities to be screened [26].

Some limitations of the study should be noted in terms of the need for further research. First, we employed the standard cut-off scores for the WHO-5 and the two-item BPHQ, whereas we optimized cut-off points for various two-item combinations of WHO-5 items in terms of their functioning in the present sample. While replication of results in another sample is warranted, we note that the consistency of findings across the two-item combinations suggests that we were not simply capitalising on chance variations in the performance of a particular combination of items in this sample. Second, like almost all previous studies of the performance of screening instruments, we focused on research diagnoses as the evaluative criteria. While a suitable comparison for many purposes, it leaves unaddressed important questions as to how the results of screening compare to unaided physician detection and how provision of the results of screening would affect physician performance. Another option would be to compare the relationship of the performance of both screening instruments and physicians to the results of formal diagnostic interviews [18, 32] with particular attention to correlates of discrepancies between results of screening and unaided physician detection. The earlier literature suggested that a considerable proportion of the patients being missed suffered from milder depression, often requiring that all of their five symptoms be detected for them to meet minimal formal diagnostic criteria with ambiguous clinical implications [7, 32]. Increased rates of physician detection [17] may have made this a more salient issue: the threshold for physicians detecting depression may have been lowered, and the group being missed may be more resistant to accepting available treatment [27]. These factors make all the more pressing the finding of a contemporary resolution of a long standing set of issues. First, does introduction of routine screening currently affect care for depression in primary care when additional support is not being provided? Furthermore, can any benefits of screening be currently sustained as physicians accumulate over time inevitable experiences with the 2/3 of positively screening patients who are not clinically depressed; with having to make judgments whether ambiguous presentations of symptoms just meet or fall short of formal diagnostic criteria; and whether to attempt to initiate treatment with the group of patients who would have remained undetected because they are adverse to accepting medication or who otherwise would not have preferred the benefits of detection and treatment over having their depression go not discussed?

Even outspoken sceptics of routine screening concede that in the context of rich resources, screening positively affects the outcome of depression and that screening can be an important part of enhancing the care of depressed patients in primary care [21]. However, the question remains of how extensive resources must be, and what are the minimal supports that sustaining screening require. Our results are consistent with suggesting that reducing screening to two questions, focusing on the absence of positive mood, and perhaps integrating a pair of such questions into routine interaction in the clinic may hold some promise.

The present study demonstrated that only two simple initial questions could be useful in detecting depression in a primary care population. Despite apparent increases in the rates of detection of depression in routine care, primary care physicians continue to miss a substantial proportion of the depressive disorders being presented to them [17]. However, with the improving rates of depression that have been observed in recent studies, a second generation task in improving the outcome of depression is ensuring adequacy of care after detection. Future work should examine the relative performance of alternative screening instruments in monitoring clinical improvement and therefore quality of care. It is likely that the forms of instruments that are optimal for screening unselected patients in the waiting room may be different for what is optimal in monitoring the quality of care for patients already in treatment [18]. All these efforts might be worthwhile, since early identification and proper treatment significantly decrease the negative impact of depression in most patients suffering from this disease [5].

Acknowledgements We want to thank the participating patients, general practitioners and practice staff. Collaborating colleagues in our research project "Recognition and Treatment of Depressive Disorders in Primary Care" within the program "German Research Network on Depression" include M. Schütze, M. D., A.-K. Allgaier, I. Seidscheck, S. Braun, S. Lösch, M. König, E. Rühl. This project is supported by the German Ministry for Education and Research within the promotional emphasis "German Research Network on Depression".

References

1. Andrews G, Peters L (1998) The psychometric properties of the Composite International Diagnostic Interview. Soc Psychiatry Psychiatr Epidemiol 33:80–88
2. Bradley CE (1996) The WHO (ten) well-being index: a critique. Psychother Psychosom 65:331–333
3. Brody DS, Hahn SR, Spitzer RL, Kroenke K, Linzer M, DeGruy FV, Williams JB (1998) Identifying patients with depression in the primary care setting: a more efficient method. Arch Intern Med 158:2469–2475
4. Chochinov HM, Wilson KG, Enns M, Lander S (1997) "Are you depressed?" Screening for depression in the terminally ill. Am J Psychiatry 154:674–676
5. Coulehan JL, Schulberg HC, Block MR, Madonia MJ, Rodriguez E (1997) Treating depressed primary care patients improves their physical, mental, and social functioning. Arch Intern Med 157:1113–1120
6. Coyne JC (2001) Depression in primary care: depressing news, exciting research opportunities. Am Psychol Soc/APS Observer 14(2) under http://www.psychologicalscience.org/observer/index.html
7. Coyne JC, Schwenk TL, Fechner-Bates S (1995) Nondetection of depression by primary care physicians reconsidered. Gen Hosp Psychiatry 17:3–12
8. Coyne JC, Thompson R, Palmer SC, Kagee A, Maunsell E (2000) Should we screen for depression? Caveats and potential pitfalls. Appl Prev Psychol 9:101–121
9. DeLong ER, DeLong DM, Clarke Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44: 837–845
10. Depression Guideline Panel (1993) Depression in Primary Care: Volume 2. Treatment of Depression, Clinical Practice Guideline, No 5. U. S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research (AHCPR Publication No. 93–0551), Rockville
11. Dunn G (2000) Statistics in Psychiatry. Arnold, London
12. Gilbody SM, House AO, Sheldon TA (2001) Routinely administered questionnaires for depression and anxiety: systematic review. BMJ 322:406–409
13. Glass RM (2003) Awareness about depression. Important for all physicians. JAMA 289:3169–3170
14. Henkel V, Mergl R, Kohnen R, Maier W, Möller H-J, Hegerl U (2003) Identifying depression in primary care: a comparison of different methods in a prospective cohort study. BMJ 326: 200–201
15. Heun R, Burkart M, Maier W, Bech P (1999) Internal and external validity of the WHO Well-Being Scale in the elderly general population. Acta Psychiatr Scand 99:171–178
16. Huppert FA, Whittington JE (2003) Evidence for the independence of positive and negative well-being: implications for quality of life assessment. Br J Health Psychol 8:107–122
17. Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR, Rush AJ, Walters EE, Wang PS (2003) The Epidemiology of Major Depressive Disorder. Results from the National Comorbidity Survey Replication (NCS-R). JAMA 289:3095–3105
18. Klinkman MS, Coyne JC, Gallo S, Schwenk TL (1997) Can case-finding instruments be used to improve physician detection of depression in primary care? Arch Fam Med 6:567–573
19. Mulrow CD, Williams JW Jr, Gerety MB, Ramirez G, Montiel OM, Kerber C (1995) Case-finding instruments for depression in primary care settings. Ann Intern Med 122:913–921
20. Murray CJ, Lopez AD (1997) Global mortality, disability, and the contribution of risk factors: Global Burden of Disease Study. Lancet 349:1436–1442
21. Palmer SC, Coyne JC (2003) Screening for depression in medical care. Pitfalls, alternatives, and revised priorities. J Psychosom Res 54:279–287
22. Paykel ES, Tylee A, Wright A, Priest RG, Rix S, Hart D (1997) The Defeat Depression Campaign: psychiatry in the public arena. Am J Psychiatry 154(6 Suppl.):59–65
23. Peters L, Clark D, Carroll F (1998) Are computerized interviews equivalent to human interviewers? CIDI-Auto versus CIDI in anxiety and depressive disorders. Psychol Med 28:893–901
24. Rogers WH, Wilson IB, Bungay KM, Cynn DJ, Adler DA (2002) Assessing the performance of a new depression screener for primary care (PC-SAD). J Clin Epidemiol 55:164–175
25. Rost K, Burnam MA, Smith GR (1993) Development of screeners for depressive disorders and substance disorder history. Med Care 3:189–200
26. Rost K, Nutting P, Smith J, Werner J, Duan NH (2001) Improving depression outcomes in community primary care practice: a randomized trial of the QUEST intervention. Quality Enhancement by Strategic Teaming. J Gen Intern Med 16:143–149
27. Rost K, Zhang M, Fortney J, Smith J, Coyne J, Smith GR Jr (1998) Persistently poor outcomes of undetected major depression in primary care. Gen Hosp Psych 20:12–20
28. Ryff CD, Singer B (1996) Psychological well-being: meaning, measurement, and implications for psychotherapy research. Psychother Psychosom 65:14–23

29. Schade CP, Jones ER Jr, Wittlin BJ (1998) A ten-year review of the validity and clinical utility of depression screening. Psychiatr Serv 49:55–61
30. Spitzer RL, Kroenke K, Williams JB (1999) Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. JAMA 282:1737–1744
31. Valenstein M, Dalack G, Blow F, Figuera S, Standiford C, Douglass A (1997) Screening for psychological illness with a combined screening and diagnostic instrument. J Gen Intern Med 12: 679–685
32. Von Korff M, Shapiro S, Burke JD, Teitlebaum M, Skinner EA, German P, Turner RW, Klein L, Burns B (1987) Anxiety and depression in a primary care clinic. Comparison of Diagnostic Interview Schedule, General Health Questionnaire, and practitioner assessments. Arch Gen Psychiatry 44:152–156
33. Whooley, MA, Avins AL, Miranda J, Browner WS (1997) Case-finding instruments for depression. Two questions are as good as many. J Gen Intern Med 12:439–445
34. Williams JW Jr, Noel PH, Cordes JA, Ramirez G, Pignone M (2002) Is this patient clinically depressed? JAMA 287:1160–1170
35. Wittchen HU (1994) Reliability and validity studies of the WHO-Composite International Diagnostic Interview (CIDI): a critical review. J Psychiatr Res 28:57–84
36. Wittchen HU, Pfister H (1997) Instruktionsmanual zur Durchführung von DIA-X Interviews. Swets Test Services, Frankfurt am Main
37. World Health Organization (1993) Composite International Diagnostic Interview, Version 1.1. World Health Organisation, Geneva
38. World Health Organization (1998) Info Package: Mastering Depression in Primary Care. World Health Organization, Regional Office for Europe, Psychiatric Research Unit, Frederiksborg